Web Based Parallel/Distributed Medical Data Mining Using Software Agents

Hillol Kargupta, Brian Stafford, and Ilker Hamzaoglu

Computational Science Methods Group
X Division, Los Alamos National Laboratory
P.O. Box 1663, MS F645, Los Alamos, NM, 87545

This paper describes an experimental parallel/distributed data mining system PADMA (PArallel Data Mining Agents) that uses software agents for local data accessing and analysis and a web based interface for interactive data visualization. It also presents results from applying PADMA for detecting patterns in unstructured texts of postmortem reports and laboratory test data for Hepatitis C patients.

Data mining involves extraction, transformation, and presentation of data in useful form. As we move more and more toward a paper-less society, each of these components of data mining is likely to face the challenges of dealing with large volumes of data and the distributed nature of the data storage and computing environments.

Medical databases are often ideal candidates for large scale, possibly distributed data mining applications. In this paper we describe an experimental software agent based system for parallel/distributed data mining. PADMA is characterized by agent based distributed data accessing, distributed data analysis, and web based interactive visualization. This paper presents results of applying PADMA in medical databases.

PADMA was used to analyze postmortem reports provided by University of New Mexico School of Medicine. Three levels of clustering (grouping) are shown in the paper. Levels indicate relations among cases with increasing generality.

Two forms of analysis are given for hepatitis C data. The first highlights body-piercing as related to blood and sexual exposures not involving IV-drug use or traditional medical exposures. The second shows the presence of tattoos as the most informative factor after accounting for the predominant exposures relating to IV drugs and transfusions.

PADMA's goal is to be a flexible system that will exploit data mining agents in parallel, for the particular application in hand. Its initial implementation used agents specializing in unstructured text document classification. PADMA agents for dealing with numeric data are currently under development. Figure 1 shows the overall architecture of PADMA. The main structural components are, (1) data mining agents, (2) facilitator for coordinating the agents, and (3) user interface.

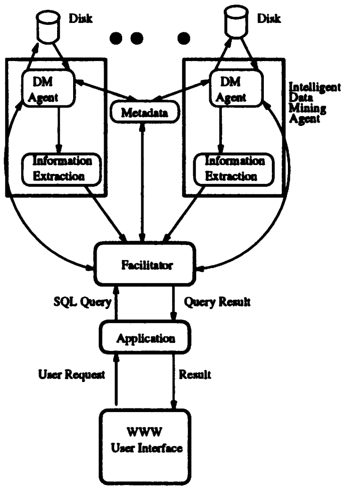

Figure 1: The PADMA architecture.

Main characteristics of PADMA are: (1) parallel query processing & data accessing, (2) parallel data analysis, (3) web based interactive data/cluster visualization. A module for supervised learning of piece-wise linear classifiers using user feedback is already developed and incorporated in PADMA. We are currently in the process of incorporating web search engines to PADMA for potential applications to web mining. For the near future, we are interested in applying PADMA to problems like disease emergence, and outbreak.

Acknowledgments

This work was supported by National Center for Supercomputing Applications and US Dept. of Energy. We also acknowledge the computing time on IBM SP2, granted by Argonne National Laboratory and the data provided by University of New Mexico School of Medicine.